Asian Journal of Applied Science and Technology (AJAST)
Volume 7, Issue 2, Pages 171-177, April-June 2023

Artificial Intelligent Decoding of Rare Words in Natural Language Translation Using Lexical Level Context and Fuzzy Semantic Reasoning

D. Samson[1*], Dr. Ignatius A Herman[2], Dr. Ibrahim Olanya Bwalya[3] & Dr. Kitraj Penchinkilas[4]

*[1]Research Scholar, Gideon Robert University, Lusaka, Zambia. [2]Director- DMI Group of Institutions, Africa. [3]Professor, Gideon Robert University, Lusaka, Zambia. [4]Department of Computer Science, Gideon Robert University, Lusaka, Zambia. Corresponding Author (D. Samson) Email: swansamson@gmail.com**

DOI: https://doi.org/10.38177/ajast.2023.7218

Article Received: 13 April 2023 *Article Accepted: 25 May 2023* *Article Published: 31 May 2023*

ABSTRACT

Artificial intelligence based Machine Translation is a Natural Language Processing, attaining significant attention to fully computerize the system that can decode basis content into the target languages. The proposed method is a Neural Machine Translation, end-to-end system, with the sentence context for converting German sentences to English sentences. The Transformer based Machine Translation is used for translation integrated with fuzzy semantic representation for handling the occurrence of rare words in the sentence level context. The FSR organization uncommon phrases collectively and constitute sentence degree context as Latent Topic Representation. WMT En-De bilingual parallel corpus is used for translation handling the Out of Vocabulary words using clustering of <UNK> tags. In the present approach there's a mismatch of translation however the proposed device is extra advanced because of the sentence context inclusion. The model performance is enhanced with hyper parameter optimization obtaining a BLEU score with a better translation of source to target language. Finally minimizing the TER score to attain a better translation rate.

Keywords: Artificial Intelligence; Neural Machine Translation; Hierarchical Phrase-based translation.

1. Introduction

This study proposes a novel neural method to enhance translation prediction by incorporating Lexical context representation. The proposed model support encodes the source language but also captures functional similarities, resulting in improved translation accuracy [1].

The model is integrated into phrase-based translation (PBT) and hierarchical phrase-based translation (HPBT) models. Extensive experiments conducted on large-scale Chinese-to-English and English-to-German translation tasks demonstrate significant improvements [2-4]. The research findings indicate that integrating the DBiCS-based neural network model (DNNJM) into the decoding process of PBT and HPBT greatly enhances the performance of statistical machine translation (SMT) [5].

The appraisal metric used is the Case Unaffected BLEU-4 score, which surpasses the performance of all baseline systems, providing superior translation quality [6]. DNNJM enthusiastically epitomizes context with various conversion time steps, leveraging structural clues for improved translation. It is worth noting that although NMT (Neural Machine Translation) still outperforms the proposed model with a significantly higher BLEU score**.**

2. Experimental Setup

The planned system is a Transformer constructed NMT with Self Commitment Technique which is added at both the Encrypt and Translator layer for translating the input language sentence (German) to the target language (English) [7]. The objective is to develop an end to end algorithm for Language Translation using NMT [8]. From the dataset 50K sentence pairs are taken and split in the proportion of 80:20 for the preparation and assessment purpose respectively.

Positional Encoding

An Absolute Positional Encoding PEj vector is added to each embedding source and target word to denote the position j in Transformer Architecture. It ensures unique output encoding for each time step, Generalize longer sentences and Deterministic in nature [9].

FSR for rare words

INPUT : Cleaned Sentence (Not in Vocabulary)

OUTPUT : Embedding vectors

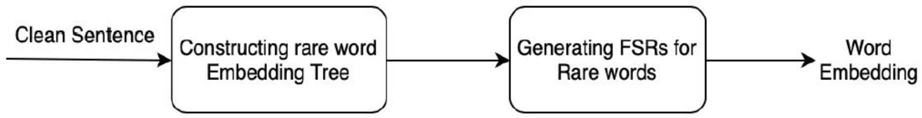

Figure 1. Fuzzy Semantic Representation

Embedding Tree Construction Algorithm

Begin

1. Train toolkit Word2Vec is used for Monolingual Data

2. Words OOV but appears >3 times are Clustered into M classes by k-means

3. Class Embedding = Mean(Word Embeddings in the class)

4. Each class has <UNKj> containing remaining words that belong to this class

(Cosine Similarity)

5. <UNKj> embedding = Mean(Remaining words in class)

End

FSR Generation Algorithm [10]

Begin

1. To address data sparseness, an input vector is constructed, this approach helps to incorporate contextual information and overcome the limitations posed by scarce data.

2. The system distinguishes between class information and word information for rare words. This differentiation allows for a more nuanced representation of rare words, considering both their inherent properties as individual words and their classification within a broader context.

3. The system computes separate input vectors for both the source and target rare words. This allows for the incorporation of contextual information specific to the input verbal and the output verbal, enabling more accurate and meaningful translations of rare words.

$$Ex_{unk} = U_x\left(\sum_j w_j Ep_j\right),\ j=1,2 \qquad (1)$$

Asian Journal of Applied Science and Technology (AJAST)
Volume 7, Issue 2, Pages 171-177, April-June 2023

$$Ey_{unk} = U_y\left(\sum_j w_j Ep_j\right),\ j=1,2 \qquad (2)$$

- The embedding of word nodes in the path, denoted as Ep2, and the embedding of class nodes in the path, denoted as Ep1, are utilized in the process. Additionally, random initialized weight matrices, Ux and Uy, are employed [12].

- The calculation or transformation of the input vectors or the overall modeling process. However, without further context or information about the specific algorithm or methodology being described, it is challenging to provide a more detailed explanation [13].

End

Latent Topic Representation [11]

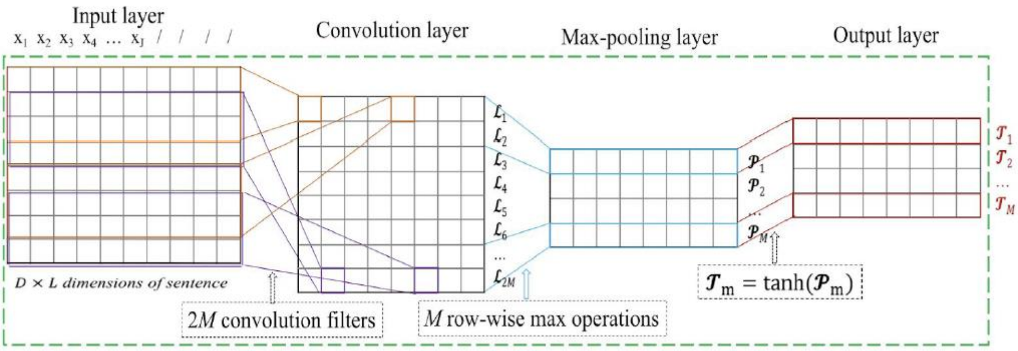

Figure 2. CNN model to represent the topic as LTR

LTR Algorithm

//Extracting Key information from the Text sentence for better Translation [14],[15]

INPUT:

{T1, T2, T3,..., Tm} is M Topics

X={ x1, x2, x3 ,..., xj } is input sentence (Word embedded)

OUTPUT:

Topic distribution of a sentence - topic context vector

Begin

1. J*D input matrix layer is passed to 2M Convolution Filters for 3 consecutive rows

where D is the dimension of the word vector and J is the sentence Length

2. The convolution Layer performs M row-wise max operations to generate Max Pooling Layer

3. Max Pooling Layer undergoes Tanh function to generate Output Layer

4. Finally T is mapped to {Key,Value} pair

End

Asian Journal of Applied Science and Technology (AJAST)
Volume 7, Issue 2, Pages 171-177, April-June 2023

3. Module Outputs

The results obtained from the Neural Machine Translation for German to English Conversion is discussed in this chapter. The preprocessing, word tokenizer and the translation sentences are added as the snapshots. The Google Translate module is integrated for the checking correct translation of the German sentences to its corresponding English sentences. The Model is trained by limiting the corpus to 50000 sentence examples. The <start> and <end> tag is attached at the end of every sentence so that the Encoder and Decoder knows its starting point and its finishing point respectively. Special characters from the sentences are removed using RE and the data gets cleaned.

Sample Datas from Corpus

```
en, de = create_dataset(path_to_file, None)
print(en[0])
print(de[0])
print()
print(en[100])
print(de[100])
print()
print(en[500])
print(de[500])
```

```
<start> go . <end>
<start> geh . <end>

<start> be kind . <end>
<start> sei nett ! <end>

<start> let s go ! <end>
<start> lass uns gehen ! <end>
```

Figure 3. After Preprocessing printing sample sentences of Parallel Corpus

Word Tokenizer

Each and every word in a sentence is mapped to a number defined in the Dictionary or vocabulary. The figure 5 shows the mapping of sample German and English sentences which is already preprocessed with its corresponding numbers tagged [5]. Using the one hot encoding in the Word2Vec model the words in the sentence are converted to its corresponding vectors. The tagging of words with indexes is helpful in maintaining the position of words in the sentence both at the foundation as fit as the target side. The Positional Encoding takes care of the position of the words in the sentence using the Trigonometric functions like Cosoidal and Sinusoidal waves.

```
Input Language; index to word mapping
1 ----> <start>
66 ----> wo
56 ----> gehen
16 ----> wir
48 ----> jetzt
244 ----> hin
7 ----> ?
2 ----> <end>

Target Language; index to word mapping
1 ----> <start>
67 ----> where
20 ----> do
17 ----> we
39 ----> go
47 ----> now
7 ----> ?
2 ----> <end>

Input Language; index to word mapping
1 ----> <start>
260 ----> gute
399 ----> nacht
17 ----> ,
53 ----> meine
7115 ----> damen
12 ----> !
2 ----> <end>
```

Figure 4. Tagging pre-processed sentences with index

ISSN: 2456-883X

Asian Journal of Applied Science and Technology (AJAST)
Volume 7, Issue 2, Pages 171-177, April-June 2023

Hyper-parameter Optimization

The hyper-parameters are used for tuning the model such as stacking the encoder and decoder layers for accurate translation thus resulting in better accuracy and maximizing BLEU score thereby minimizing TER score. The parameters which affect the system are buffer_size, batch_size, embedding_dimensions, units, vocab_inp_size and vocab_tar_size.

```
BUFFER_SIZE = len(input_tensor_train)
BATCH_SIZE = 64
steps_per_epoch = len(input_tensor_train)//BATCH_SIZE
embedding_dim = 128
units = 1024
vocab_inp_size = len(inp_lang.word_index)+1
vocab_tar_size = len(targ_lang.word_index)+1
```

Figure 5. Hyper-parameters for Transformer Architecture

Sentence Translation

Thus Transformer based NMT is used for translating German to English sentences where the results show its a close approximation to the Google Translator. The Attention Graph is plotted for the translation where the sentence context is used to get better results with the support of Attention weights. The brighter color in the graph is the closest translation to its corresponding word of a sentence.

Test cases

A check case is a file which has a hard and fast of situations or movements that are finished at the proposed gadget software with a purpose to affirm the predicted capability of the feature. The Sentences which are translated correctly and with close approximation based on the word of the sentence and the rare words are handled.

Inference

From the sample test cases below the actual output obtained by the Google translator API for the German sentences is taken and compared with the predicted output given by our system. The graph plotted is based on the attention weight of the German sentence with respect to its English sentence. The brightest region in the attention graph tells the correct translation. In most of the cases there is a close relationship with the Google Translator output. Due to the sentence level context, the translated output shows similar meaningful words with the actual output. Rare words in the sentence are translated based on context meaning based on attention weight mechanism with Key, Query and Values used in the transformer model. The Encoder-Decoder model takes all the context as LTR and the FSR for OOV rare words is taken into consideration for a better translation output for German to English sentences. 9/10 test cases passed by the translator system.

Training Time

The total time taken by the batches per epoch is tabulated below. Each epoch has 6 batches of size 100 to be trained and the sum of all the time of batches per epoch is calculated. The GPU with RAM of 25 GB is used to run the model and the time taken per epoch is given in seconds (sec). The training time for every 5 epochs increases by a slight margin.

Asian Journal of Applied Science and Technology (AJAST)
Volume 7, Issue 2, Pages 171-177, April-June 2023

Table 1. Training time for epochs

EPOCHS VS TIME	
EPOCHS	TIME (in sec)
1-5	189.2447
6-10	193.5873
11-15	193.3618
16-20	193.8892
21-25	194.2428

4. Conclusion

In this Research, we develop an end-to-end Language Translation structure that utilizes a Transformer-based (NMT) to convert German sentences into English sentences. The system takes the context of a sentence into consideration to ensure that the translation internments the same sense as the basis verdict. The Transformer's Self-Courtesy contrivance is employed to increase the accuracy of the translated verdicts. To handle rare words that are out-of-vocabulary (OOV) in the sentence, we integrate Frequency Sensitive Replacers (FSR) with the Transformer model. Additionally, we incorporate a (CNN) within the system to leverage the context of the sentence as the Language Translation Resource (LTR). The process of modeling the Sequence-to-Sequence system for the dataset is time-consuming. However, the experiments demonstrate that our proposed NMT system achieves excellent performance, as evidenced by a higher BLEU score compared to conventional Machine Translation (MT) systems. Furthermore, the system significantly reduces the Translation Error Rate (TER).

Declarations

Source of Funding

This study did not receive any grant from funding agencies in the public or not-for-profit sectors.

Competing Interests Statement

Authors have declared no competing interests.

Consent for Publication

The authors declare that they consented to the publication of this study.

Authors' Contributions

All authors equally participated in research and drafting the manuscript.

References

[1] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). Neural machine translation by jointly learning to align and translate. ArXiv Preprint ArXiv: 1409.0473.

[2] Cao, Ziqiang, et al. (2015). A novel neural topic model and its supervised extension. Twenty-Ninth AAAI Conference on Artificial Intelligence.

[3] Suresh Kumar, K., Helen Sulochana, C., Radhamani, A.S., Ananth Kumar, T. (2022). Sentiment Lexicon for Cross-domain Adaptation with Multi-domain Dataset in Indian Languages Enhanced with BERT Classification Model. Journal of Intelligent & Fuzzy Systems, 43(5): 6433-6450. doi: 10.3233/JIFS-220448.

[4] Chen W., Matusov E., Khadivi S., Peter J. (2016). Guided alignment training for topic-aware neural machine translation. In 38th Annual Meeting & Symposium of the Antenna Measurement Techniques Association, Volume 1, Austin, Texas, November 2016, Pages 121–134.

[5] Chitnis R. and DeNero J. (2015). Variable-length word encodings for neural translation models. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Pages 2088–2093.

[6] Chung J., Cho K., and Bengio Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. In Proceedings of the 54th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Pages 1693–1703.

[7] Fadaee M., Bisazza A., and Monz C. (2017). Learning topic-sensitive word representations. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada: Association for Computational Linguistics, July 2017, Pages 441–447.

[8] Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao (2020). Unsupervised Neural Machine Translation With Cross-Lingual Language Representation Agreement. IEEE/ACM Trans on Audio, Speech, and Language Processing, Volume 28.

[9] Kim, T.Y., Bae, S.H., & An, Y.E. (2020). Design of smart home implementation within IoT natural language interface. IEEE Access: Practical Innovations, Open Solutions, 8: 84929–84949.

[10] Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao (2019). Neural Machine Translation with Sentence-level Topic Context. IEEE/ACM Trans on Audio, Speech, and Language Processing. doi: https://doi.org/10.1109/TASLP.2019.2937190.

[11] Kim Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar: Association for Computational Linguistics, October 2014, Pages 1746–1751.

[12] Muyun Yang, Shujie Liu, Kehai Chen, Hongyang Zhang, Enbo Zhao, and Tiejun Zhao (2020). A Hierarchical Clustering Approach to Fuzzy Semantic Representation of Rare Words in Neural Machine Translation. IEEE Trans. on Fuzzy Logic, 28(5).

[13] Rui Wang, Masao Utiyama, Andrew Finch, Lemao Liu, Kehai Chen, and Eiichiro Sumita (2018). Sentence Selection and Weighting for Neural Machine Translation Domain Adaptation. IEEE/ACM Trans on Audio, Speech, and Language Processing. doi: https://doi.org/10.1109/TASLP.2018.2837223.

[14] Snover, Matthew, et al. (2006). A study of translation edit rate with targeted human annotation. Proceedings of association for machine translation in the Americas, 200(6).

[15] Xing Wang, Zhaopeng Tu, Min Zhang (2015). Incorporating Statistical Machine Translation Word Knowledge into Neural Machine Translation. Journal of Latex Class Files, 14(8).